Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 18 (2013) 2396 – 2405

Procedia Computer Science

International Conference on Computational Science, ICCS 2013

Identification and visualization of dominant patterns and anomalies in remotely sensed vegetation phenology using a parallel tool for principal components analysis

Richard Tran Mills[a,b,*], Jitendra Kumar[a], Forrest M. Hoffman[a,c], William W. Hargrove[d], Joseph P. Spruce[e], Steven P. Norman[d]

[a]*Oak Ridge National Laboratory, Oak Ridge, TN, USA*
[b]*University of Tennessee, Knoxville, TN, USA*
[c]*University of California, Irvine, CA, USA*
[d]*Eastern Forest Environmental Threat Assessment Center (EFETAC), Southern Research Station, USDA Forest Service, Asheville, NC, USA*
[e]*NASA Stennis Space Center, Bay St. Louis, MS, USA*

Abstract

We investigated the use of principal components analysis (PCA) to visualize dominant patterns and identify anomalies in a multi-year land surface phenology data set (231 m × 231 m normalized difference vegetation index (NDVI) values derived from the Moderate Resolution Imaging Spectroradiometer (MODIS)) used for detecting threats to forest health in the conterminous United States (CONUS). Our goal is to find ways that PCA can be used with this massive data set to automate the process of detecting forest disturbance and attributing it to particular agents. We briefly describe the parallel computational approaches we used to make PCA feasible, and present some examples in which we have used it to visualize the seasonal vegetation phenology for the CONUS and to detect areas where anomalous NDVI traces suggest potential threats to forest health.

Keywords: phenology; MODIS; NDVI; remote sensing; principal components analysis; singular value decomposition; data mining; anomaly detection; high performance computing; parallel computing

1. Introduction

Early identification of forested areas threatened by insects, disease, drought, or other agents can be critical to preventing long-term or irreversible damage to forest ecosystems. With well over 600 million acres of forest and wildlands in the United States, however, it is impossible for federal and state agencies to regularly monitor any significant fraction of these lands through aerial surveys and ground-based inspections, causing many threats to go unnoticed until it is too late to easily take action to mitigate or correct them. To address this problem, a collaboration led by the USDA Forest Service, in cooperation with NASA Stennis Space Center, USGS, and the Department of Energy Oak Ridge National Laboratory, has been been developing *ForWarn*, a forest change recognition and tracking system that uses high-frequency, moderate resolution satellite data to monitor changes in forest cover and health. *ForWarn* comprises the first tier of a broader National Early Warning System for Forest Health Threats, which helps direct higher-resolution, second-tier aerial and ground-based surveys.

*Corresponding author. Tel.: +1-865-241-3198
E-mail address: rmills@ornl.gov, rtm@utk.edu.

Selection and peer review under responsibility of the organizers of the 2013 International Conference on Computational Science
doi:10.1016/j.procs.2013.05.411

ForWarn identifies changes in forest state through analysis of satellite-derived data indicating the temporal variation of vegetation "greenness", i.e., the ecosystem phenology. Specifically, it uses Normalized Difference Vegetation Index (NDVI) values derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) sensors aboard NASA's Aqua and Terra satellites as a proxy for vegetation phenology. Leaf cells strongly absorb solar radiation in the photo synthetically active radiation range (wavelengths of 400 to 700 nm) and scatter radiation in the near-infrared spectral region (wavelengths > 700 m). NDVI exploits these differences in reflectances to provide a measure of vegetation canopy "greenness":

$$\mathrm{NDVI} = \frac{\left(\sigma_{\mathrm{nir}} - \sigma_{\mathrm{red}}\right)}{\left(\sigma_{\mathrm{nir}} + \sigma_{\mathrm{red}}\right)} \tag{1}$$

These spectral reflectances are ratios of reflected over incoming radiation, $\sigma = I_r/I_i$, hence they take on values between 0.0 and 1.0. As a result, NDVI varies between −1.0 and +1.0. Dense vegetation cover is 0.3–0.8, soils are about 0.1–0.2, surface water is near 0.0, and clouds and snow are negative. *ForWarn* utilizes NDVI values derived from the 231 m × 231 m resolution MOD 13 Gridded Vegetation Indices products from the Aqua and Terra satellites, which are generated every 16 days; because the Aqua and Terra products are staggered in time, a new MOD 13 product is available every 8 days. These products are further processed by NASA Stennis Space Center into a smoothed and quality-controlled data set that provides 46 NDVI values per year at every location.

ForWarn is currently providing interactive forest disturbance detection maps though the U.S. Forest Change Assessment Viewer website.[1] These maps are computed through raster map arithmetic approaches in which current NDVI values are compared with values from some historical baseline. E.g., the current maximum NDVI observed over a 24-day window at a given location may be compared with the maximum NDVI for the same location/window observed over a set of previous years. Such approaches have demonstrated their utility at identifying several types of forest disturbances [1], but a potential difficulty is identification of appropriate parameters (maximum NDVI, 20% "spring" NDVI, window size, years to use in establishing the historical baseline, etc.). This has prompted us to explore geospatiotemporal data-mining techniques that use high-performance computing to analyze the entire MODIS-based NDVI history for the entire contiguous United States (CONUS) to establish a potential basis for determining what constitutes "normal" seasonal and inter-seasonal variation expected at a given geographic location, and for determining what constitutes a departure from "normal" that is significant enough to merit further scrutiny.

One set of approaches we have explored are based on k-means cluster analysis of the entire NDVI history for the CONUS. These approaches are described in more detail in [2, 3], but we summarize them here. For each year and each grid cell in the CONUS, the NDVI values are arranged into an observation vector of 46 NDVI values representing the seasonal NDVI trace for that year/location. All of these observation vectors are combined into a data matrix with 46 columns and hundreds of millions of rows (each year corresponds to 146.4 million rows). The data are then standardized (the column mean is subtracted from each element of a column, and each element is divided by the column standard deviation) and then clustered using a highly-parallel k-means clustering code [4]. The resulting cluster assignments are then mapped back to each map cell and year from which each observation came, yielding one map per year in which each cell is classified into one of k clusters or "phenoclasses", which can be viewed as forming a dictionary of prototypical annual NDVI traces that are derived from the full spatiotemporal extent of the observations comprising the input data set. Disturbance or recovery can be detected by analyzing the history of phenoclass assignment in a variety of ways, such as looking for significant deviation from the statistical mode of cluster assignments or looking for a large Euclidean distance between the currently assigned cluster centroid and those from a prior year or years (which we have termed the *transition distance*).

Our experiments with these clustering-based approaches have shown that they are effective at identifying a wide range of disturbances, particularly those involving high mortality events such as fire, storms, or mountain pine beetle outbreaks. Identifying slower-acting agents, such as hemlock woolly adelgid, that cause a gradual decline in forest health is more difficult. Also, the annual phenology of some areas is highly influenced by interannual climate variability: grasslands, for instance, experience rapid greenup after precipitation and do not

[1] http://forwarn.forestthreats.org/fcav

have smooth annual cycles. These areas tend to display a large transition distance from year to year even when there is essentially no real change in the vegetation health. To remedy these shortcomings of our cluster analysis-based approaches, we have been exploring the use of principal components analysis (PCA) to complement these approaches. In this paper we present some of the initial PCA-based approaches we have tried and the results of some of our experiments with them.

2. Principal Components Analysis

PCA can be regarded as a procedure for computing the most meaningful or natural basis for expressing a data set. For a given n-dimensional data set, it determines a set of orthonormal basis vectors or axes (the *principal components*) that are a linear combination of the original basis and possesses the following property: the first axis explains the greatest amount of the variance of the data set as possible on a single axis, while the second explains the next largest portion of the variance possible, and so forth. Formally, the first principal component $\mathbf{v}_1$ points in the direction of maximum variance in the (mean-centered) $m \times n$ data matrix $\mathbf{X}$:

$$\mathbf{v}_1 = \arg\max_{\|\mathbf{v}\|=1} \mathrm{Var}(\mathbf{X}\mathbf{v}). \tag{2}$$

Because $\mathrm{Var}(\mathbf{X}) = \|\mathbf{X}\|^2$ if $\mathbf{X}$ is mean-centered, this is equivalent to

$$\mathbf{v}_1 = \arg\max_{\|\mathbf{v}\|=1} \|\mathbf{X}\mathbf{v}\|. \tag{3}$$

Given any $k-1$ principal components, the kth principal component maximizes the variance of the "residual" data that do not map onto the first $k-1$ components:

$$\mathbf{v}_k = \arg\max_{\|\mathbf{v}\|=1} \|(\mathbf{X} - \sum_{i=1}^{k-1} \mathbf{X}\mathbf{v}_i\mathbf{v}_i^T)\mathbf{v}\| \tag{4}$$

The principal components can be determined by computing the eigenvectors and eigenvalues of the covariance matrix (or correlation matrix, if the data are standardized, as in the case of our NDVI dataset) $\mathbf{C}_\mathbf{X} = \frac{1}{m-1}\mathbf{X}^T\mathbf{X}$. (For an excellent and highly intuitive explanation of why, see [5].) Ordering the eigenvectors $\mathbf{v}_i$ by magnitude of their corresponding eigenvalues λ_i, these eigenvectors comprise the principal components. The projection of the data matrix onto the ith principal component is given by $\mathbf{X}\mathbf{v}_i$, and the proportion of variance explained on this axis is $\lambda_i / \sum_{j=1}^{n} \lambda_j$ (which is equivalent to λ_i/n in the case of a standardized data matrix).

The construction of the principal components allows one to easily examine the intrinsic dimensionality of a data set, and PCA is routinely used for dimensionality reduction and to extract the dominant trends in a data set. If a set of the first p components explains most of the variance, the data can be largely represented using this reduced basis. Recalling that the principal component vectors are formed from linear combinations of the original variables, the individual elements (that is, the coefficients of the linear combination) of the most significant principal components can be examined to determine the relative contributions of the original variables to the total variance of the data set. In section 4 we illustrate how PCA can help us understand and visualize some of the dominant trends in our national phenology data set.

Although PCA is commonly employed to extract the dominant structure or trends of a data set, it also has utility in identifying anomalous features of the data set. If the first p components that together explain most of the variance can be considered as a vector basis for the subspace of "normal" behavior, the $n-p$ remaining vectors can be considered to span the subspace of "abnormal" behavior. In section 5 we explore how these notions can be used to separate NDVI traces into "normal" and "abnormal" components.

3. Computing the PCA for Large Data Sets

We have stated that PCA can be performed by computing the eigenpairs of the covariance matrix of the mean-centered data matrix $\mathbf{X}$. In practice, computing the PCA in this manner may be undesirable because of the

round-off errors introduced in forming the matrix-matrix product $\mathbf{X}^T\mathbf{X}$. Instead, the PCA is usually computed by forming the singular value decomposition (SVD)

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{5}$$

of the matrix $\mathbf{A} = \frac{1}{\sqrt{n-1}}\mathbf{X}$, which is related to the covariance matrix of the data set by $\mathbf{A}^T\mathbf{A} = \mathbf{C}_\mathbf{X}$. The columns of $\mathbf{V}$ (the right singular vectors) form the principal components of $\mathbf{X}$ and the non-zero entries of the diagonal matrix $\mathbf{\Sigma}$ are the square roots of the eigenvalues of $\mathbf{C}_\mathbf{X}$.

Our national phenology data set is very large: each year consists of 46 NDVI values at each of 144.6 million grid points. In single precision, each year of data is about 27 GB in size, so the size of the data set for the entire 12 years of the MODIS NDVI record is 324 GB. Working with even a fraction of this data set requires distributed memory parallelism due to both wall-clock time and memory constraints, so we have constructed a Message Passing Interface (MPI)-based code to perform the PCA and data projection. Our (dense) matrices and vectors are distributed across a logically two-dimensional MPI process grid (our code supports a 2D grid, but for the matrix dimensions here we choose a processor grid that consists of only one column) using a standard block-cyclic partitioning, and routines from the PLAPACK [6] library are used to perform linear algebra operations on these objects.

PLAPACK includes a routine, `PLA_SVD`, that calculates the thin SVD by first reducing the matrix $\mathbf{A}$ to bidiagonal form ($\mathbf{A} = \mathbf{U}_1\mathbf{B}\mathbf{V}_1^T$) through a series of unitary Householder reflections, and then computing the SVD of the bidiagonal matrix $\mathbf{B} = \mathbf{U}_2\mathbf{\Sigma}\mathbf{V}_2^T$. The singular vectors of $\mathbf{A}$ are then $\mathbf{U} = \mathbf{U}_1\mathbf{U}_2$ and $\mathbf{V} = \mathbf{V}_1\mathbf{V}_2$. The reduction to bidiagonal form happens in parallel, and the SVD computation of the (very small) bidiagonal matrix is performed sequentially via the `DBDSQR` routine in LAPACK. We originally used the `PLA_SVD` routine, before realizing that because our matrices are extremely "tall and skinny" (the number of rows m is in the millions while then number of columns n is only 46), it is significantly more efficient to use the Lawson-Hanson-Chan algorithm: the Golub-Kahan algorithm used by `PLA_SVD` requires approximately $4mn^2 - \frac{4}{3}n^3$ operations, whereas Lawson-Hanson-Chan requires approximately $2mn^2 + 2n^3$; the latter also has the additional advantage of being based on QR factorization, which is a one-sided operation and therefore has better parallel efficiency than the two-sided bidiagonalization process. With this approach, we do the following:

1. Form the reduced QR factorization $\mathbf{A} = \mathbf{QR}$ ($\mathbf{Q}$ is orthonormal and $\mathbf{R}$ is upper-triangular); this is done in parallel via the `PLA_QR` routine that uses a sophisticated block (matrix-matrix multiply) based algorithm [7].
2. Perform an `MPI_Gather` operation to gather the $n \times n$ matrix $\mathbf{R}$ to process 0.
3. Process 0 calls the serial LAPACK algorithm `DGESVD` to compute the SVD $\mathbf{R} = \mathbf{U}\mathbf{S}\mathbf{V}^T$. The matrices $\mathbf{S}$ and $\mathbf{V}$ then hold the relevant portions of the SVD of $\mathbf{A}$. (The left singular vectors can also be computed by forming $\mathbf{QU}$, but we do not need them for our PCA.)
4. Process 0 scatters $\mathbf{V}$ to the other processes according to the block-cyclic distribution.
5. If projection of the original data onto the principal component space is desired, call the `PLA_Gemm` routine to perform the parallel matrix-matrix multiply.

There is a serial bottleneck in the code where the SVD of $\mathbf{R}$ is computed, but this matrix is so small (only 46×46 for our NDVI data set) that this serial portion is essentially negligible. We note that the parallel efficiency of our code could be increased by using the "communication-avoiding" Tall Skinny QR factorization algorithm presented in [8].

4. Visualizing Dominant Patterns: Similarity Color Maps

The first few principal components provide a useful means to visualize the dominant patterns in the yearly phenoclass-assignment maps generated by our clustering tool. Taking the first three components (which together typically explain almost 95% of the variance), we apply a varimax rotation, which orthogonally rotates the components such that each will have large loadings on only a few of the original variables. Figure 1 displays the loadings for the rotated components derived from a principal components analysis of all of the $k = 1000$ cluster centroids for years 2000–2010 (we note that for our purpose here, working with the PCA of the full data set is not necessary, as the maps we will generate quickly converge to the same visual appearance as k increases). One can see that the three components correspond roughly to the first third, middle third, and last third of the year, respectively. We

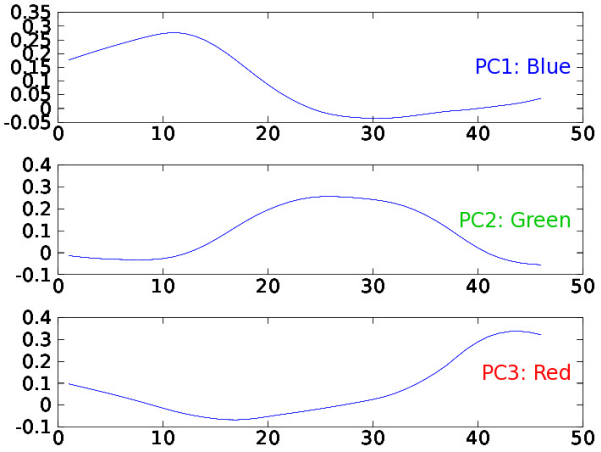

Fig. 1. The loadings (coefficients in the linear combination of the 46 original variables) along the three varimax-rotated principal axes. The x-axis corresponds to the eight-day NDVI-acquisition windows and loadings are shown on the y-axis.

can then use these three components to construct maps of the yearly phenoclass assignments (cluster membership) that use *similarity colors* [9] instead of random colors to indicate the cluster assignments: each cluster center is projected onto the three rotated axes and then these scores are mapped onto the three channels of the red-green-blue color space. The similarity of the annual phenology of any two points on the map can then be compared by a simple visual inspection of how closely they resemble each other in color.

Figure 2 depicts a similarity-colored map for the CONUS in year 2000 clustered at the $k = 1000$ level of division. The first varimax-rotated component (which we will refer to as *PC1*) has been assigned to blue, the second (*PC2*) to green, and the third (*PC3*) to red. Several patterns can be discerned on this map. The agriculture dominated (large amplitude, highly productive deciduous vegetation) in the Midwest and the highly irrigated Central Valley of California are the greenest parts of the map. This makes sense because both display very high NDVI during the growing season and peak just after mid-summer, so they score very highly onto PC2. The Central Valley is a paler green, indicating that it scores higher on PC1 and PC3 than the Midwest, likely because some crops are grown there year-round. The highly productive evergreen biomes in the Pacific Northwest and the Gulf Coast are purple because they have the highest NDVI outside of the growing season. The peaks of the Sierra Nevada are an even darker purple, perhaps because they are snow-covered part of the year and their productivity is more limited. The evergreens present in northern portions of Minnesota, Wisconsin, and Michigan also show up in purple, as do the evergreen forests of Maine. The purple strip of ponderosa pine in central Arizona can also be seen. Much of the Appalachians and Northeast are aquamarine because they score highly on PC1 and PC2, with streaks of purple on ridgetops. These are mixed deciduous (PC2) and evergreen (PC1 + PC3) systems. Certain swamps, wetlands, and lakes are dark olive drab because they have a constant and very low NDVI. Similarly, deserts are dark yellow because they have constant and very low NDVI. Interestingly, urban areas are also somewhat purple, probably because of managed vegetation that may be green year-round. We are unsure about why some canyon lands show up as almost pure red to pink, except that absolutely nothing grows and PC3 is lowest for the longest part of the year.

5. Principal Components for Anomaly Detection

Although PCA is most commonly employed to extract the dominant structure and trends of a data set, it can also be useful for identifying anomalous features. If most of the variance of a data set can be explained using the first p principal components, observations that fit the general patterns of the data will be well-represented in this subspace and will score highly along these components. Observations that do not fit the normal correlation structure, conversely, will have strong scores along the lower-order $n - p$ components. Methods based on this property of the PCA subspaces have been used for automated detection of anomalous traffic in computer networks [10, 11].

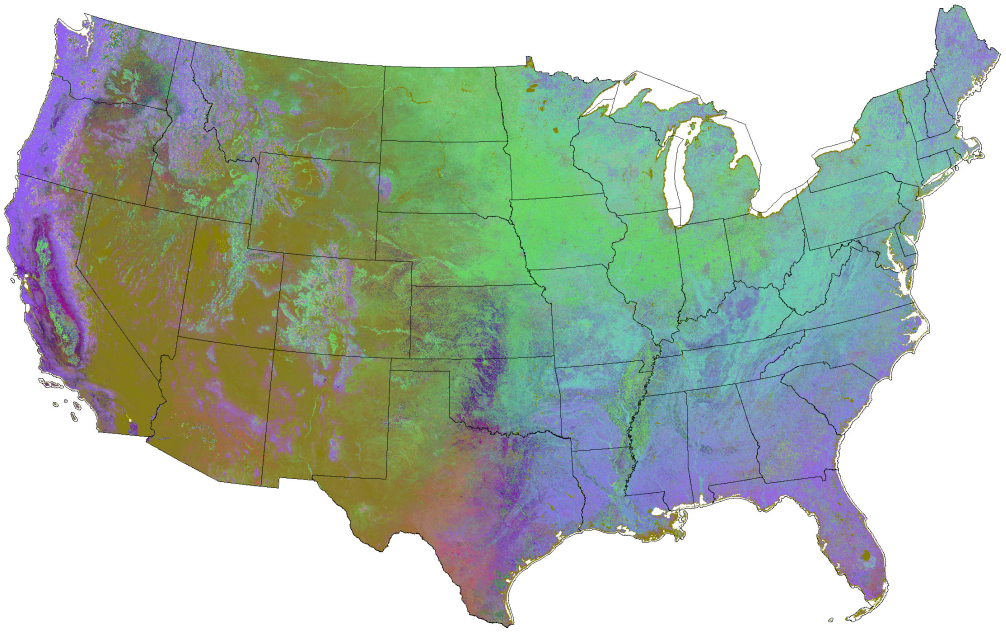

Fig. 2. Phenoclass assignment map for year 2000 with $k = 1000$. Similarity colors are used to indicate cluster membership.

Here, we experiment with using the lower-order principal components to identify anomalous features within our NDVI data set (possibly indicative of disturbance or recovery) via the following simple procedure: For a given observation vector $\mathbf{x}$, determine its representation $\mathbf{y} = [y_1, ..., y_n]$ (that is, its *scores*) in terms of the principal components basis. Then calculate the function

$$f(\mathbf{y}) = \sum_{i=p-n+1}^{p} \frac{y_i^2}{\lambda_i}, \tag{6}$$

which we note is equivalent to the Mahalanobis distance from the origin in the space formed by the lower-order principal components. If $f(\mathbf{y})$ exceeds some threshold, then the observation vector is flagged as anomalous. We note that under the assumption of normality and for sufficiently large sample size, $f(\mathbf{y})$ follows a chi-squared distribution, so this threshold could be used to identify outliers. However, here we have chosen this threshold based on a user-specified quantile of the empirical distribution of the function.

It is important to note that this approach will flag any observations that are somehow "unusual" *for the collection of data from which the principal components have been calculated.* Thus the choice of the spatiotemporal subset of the NDVI dataset fed as input to the PCA calculation will affect what constitutes a "normal" or "abnormal" observation. If, say, the entire NDVI data set is used in the PCA, a heavily disturbed region in Florida with very low productivity may be classified as "normal" because it is very similar to many observations from the Mojave Desert that are present in the data set; whereas it might be classified as abnormal using a set of principal components computed from a subset of observations only from the humid Southeast. Clearly, the choice of both temporal and spatial window over which the PCA is computed requires some judgement. We plan further exploration of this issue, but for the initial experiments described here we have chosen to work with PCAs computed over single years and within a spatial domain conforming to the eco-climatic domains established by the National Ecological Observatory Network (NEON). The NEON domains have been quantitatively delineated through multivariate geographic clustering of a national data set of eco-climatological variables [12, 13]; we choose to work with those here because some degree of eco-climatological (and hence phenological) homogeneity can be expected

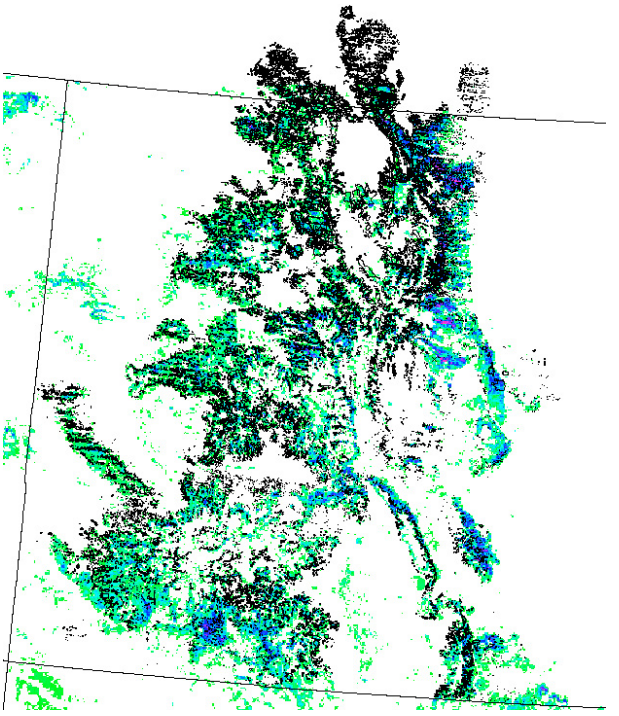

Fig. 3. A portion of the PCA-based anomaly map (maps cells scoring in the 85th percentile are shown) for the Southern Rockies–Colorado Plateau NEON Domain for year 2008, showing flagged anomalies in Colorado and southern Wyoming. Aerial detection survey perimeters are shown in black.

within each domain. We present several examples below in which the PCA-based anomaly detection approach appears to have utility in identifying forest disturbances. In all of the examples, principal component vectors 10–46 are used as the basis for the "abnormal" space, which explains 5–10% of the variance in each domain. The threshold for $f(\mathbf{y})$ used to flag observations as anomalous is chosen such that only those observations that fall into a high percentile (we varied this for each domain) are flagged. In all of the examples, certain features that are not disturbances but possess very anomalous NDVI traces (e.g., bodies of water) show up very strongly.

Figure 3 depicts a portion of the PCA-based anomaly map computed for the Southern Rockies–Colorado Plateau NEON Domain for year 2008. Any map cells that are colored-in have a Mahalanobis distance $f(\mathbf{y})$ that is in the 85th percentile. (This map, as well as the maps in Figures 4 and 6, is mostly intended as a binary map—that is, a map cell is either flagged as anomalous or not—but the anomalous cells are colored by score from low to high in the order blue, green, yellow, and then red, as this aids in visual identification of features.) Polygons outlined in black enclose areas flagged as damaged by aerial detection surveys (ADS). Forests in the region depicted have been extensively damaged by a widespread mountain pine beetle outbreak that began in 2003 and has killed very large numbers of ponderosa and lodgepole pines, as well as by both long-term and sudden aspen decline. We observe very high correspondence between the areas we have flagged as anomalous and the ADS polygons (which we note are somewhat inexact by nature).

We have also examined the Southeast NEON Domain and found that hurricane-induced mortality can easily be discerned in the vicinity of the Louisiana coast. Figure 4 depicts PCA-based anomaly maps for the years 2004–2009, and Figure 5 depicts the entire NDVI history for a near-coastal location in southwestern Louisiana. This area was strongly affected by Hurricanes Katrina and Rita in 2005, and later by Hurricanes Ike and Gustav in 2008. Strong anomalies appear in 2005, which decrease in 2006 as some recovery occurs. The 2007 map looks fairly normal, and then strong anomalies again reappear in 2008. The appearance and disappearance of these strong anomalies on the map correspond well with what one would expect to see based on the example NDVI trace.

Lastly, we present in Figure 6 a portion of the PCA-based anomaly map for the year 2010 in the Southern Appalachians/Cumberland Plateau NEON Domain, centered roughly around the high-elevation border between

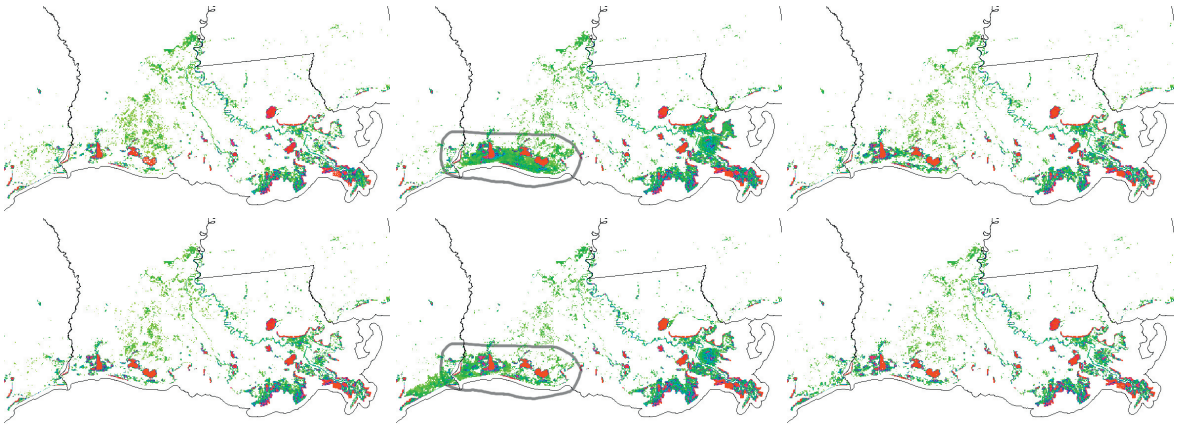

Fig. 4. Portions of the PCA-based anomaly maps (map cells scoring in the 90th percentile are shown) for the Southeast NEON Domain for years 2004–2009, showing the area in the vicinity of the Louisiana coast. From left to right, the top row shows years 2004, 2005, and 2006, respectively, and the bottom row years 2007, 2008, and 2009. The affected regions are circled in the 2005 and 2008 maps. The prominent red features are water bodies.

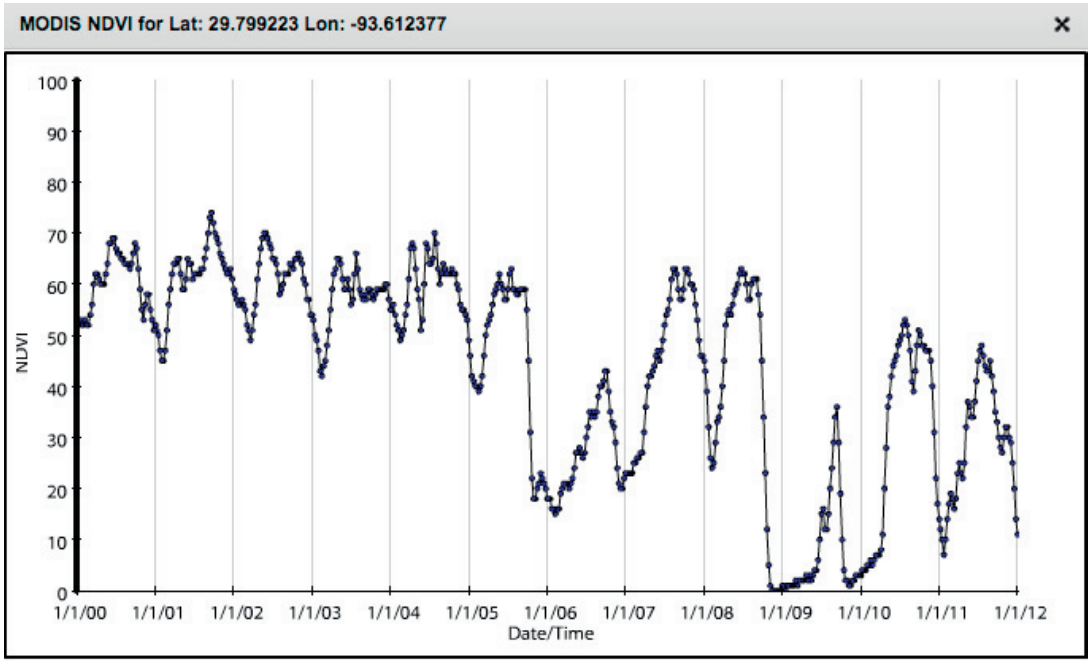

Fig. 5. NDVI trajectory as viewed via the Forest Change Assessment Viewer for a location (close to the center of the circled region in Figure 4) near the coast in southwestern Louisiana showing apparent hurricane-induced mortality from events in 2005 and 2008.

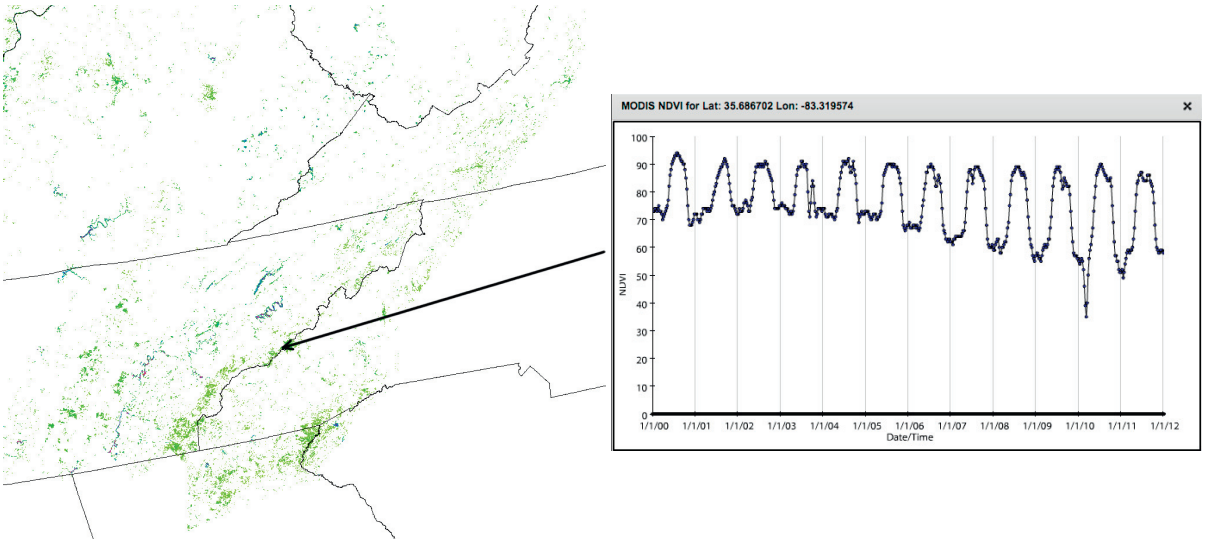

Fig. 6. At left, a portion of the PCA-based anomaly map (map cells scoring in the 90th percentile are shown) for the Southern Appalachians/Cumberland Plateau NEON Domain for year 2010. A variety of "unusual" features are shown in this map, including rivers, but many are as yet unexplained, though potentially due to multiple factors. The arrow indicates a location thought to be affected by hemlock woolly adelgid, and the corresponding NDVI trajectory as viewed via the Forest Change Assessment Viewer is shown at right. The decline in the seasonal minimum with a smaller decline in the seasonal maximum is consistent with a shift in the dominance of evergreen toward deciduous vegetation and an overall reduction in greenness due to a lower leaf density, respectively. The sudden drop in NDVI in 2010 is due to multiple weeks of snow cover, an unusual event that prevents accurate remote sensing of snow-free vegetated NDVI.

Tennessee and North Carolina that runs along the crest of the Unaka Range. Several anomalous regions show up on the map, mostly in coniferous forests in higher elevations. Attribution of these anomalies to any particular agent is challenging, however. The health of these coniferous forests has been significantly affected by several factors, including acid rain, balsam woolly adelgid, and a progressive and serious problem with hemlock woolly adelgid. We are particularly interested in identifying areas affected by hemlock woolly adelgid, which causes decline spanning a number of years and may be difficult to detect with our clustering-based methods. The right side of Figure 6 shows an example NDVI trace from an area that may be affected by the hemlock woolly adelgid. Such areas show up well in the anomaly maps, but most evergreen forests at higher elevations also show up. Because these forests have been widely affected by various destructive agents, it is hard to say whether these show up due to disturbance or simply due to factors such as the highly-variable mountain climate. Considerably more investigation will be required to determine if this PCA-based technique can be useful in these regions. It may be that selecting the Southern Appalachians NEON Domain as the spatial window for areas included in the PCA is a bad choice because the high-elevation areas make up a relatively small portion; we may need to use a smaller window that has a greater proportion of high-elevation areas. It may also be the case that such unsupervised PCA-based techniques are not useful on their own for identifying these agents; they may need to be used in conjunction with clustering-based techniques or supervised classification methods.

6. Conclusions and Future Directions

We developed a parallel tool for performing principal components analysis on large (dense) data sets and used it to inspect both dominant patterns and anomalous observations in our NDVI data set. Using the first three principal components to construct a table of similarity colors greatly aids interpretation of the geographic patterns in seasonal vegetative phenology observed in the data set. Using the subspace of very low-order principal components to identify anomalous observations that correspond to forest disturbance appears promising: without using any historical baseline, some types of disturbance—including some that are may not be high-mortality events—show up very well. Considerably more experimentation with and validation and refinement of this approach is required, however. In our experiments thus far we have used only observations from the same NEON domain and

the same year as input to the PCA tool; we clearly need to experiment with using both larger and smaller spatial and temporal windows for selecting the input data set. Analyzing the entire data set will require some updates and optimization of the parallel PCA tool: the data set has become large enough that updating the tool to support 64-bit indexing is necessary, and support for parallel I/O (the data are currently simply distributed by process 0) is needed as well. Beyond experimenting with the PCA-based technique on its own, we also need to explore its use in conjunction with the raster map arithmetic and clustering-based methods we have developed. Perhaps the most interesting work is to explore the use of supervised PCA-based approaches: if a "dictionary" of observations corresponding to many different types of disturbances can be compiled, we might be able to classify new observations as belonging to a particular disturbance class by seeing how well they can be represented using the principal components basis generated by PCA of the observations forming each disturbance class. Existing methods being used in the *ForWarn* system are already fairly effective at detecting disturbance, but attribution is much harder and more time-consuming, so such an approach could prove very useful.

7. Acknowledgments

This research was sponsored by the U.S. Department of Agriculture Forest Service, Eastern Forest Environmental Threat Assessment Center. This research used resources of the National Center for Computational Science at Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC, for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

8. References

[1] W. W. Hargrove, J. P. Spruce, G. E. Gasser, F. M. Hoffman, Toward a national early warning system for forest disturbances using remotely sensed phenology, Photogramm. Eng. Rem. Sens. 75 (10) (2009) 1150–1156.

[2] F. M. Hoffman, R. T. Mills, J. Kumar, S. S. Vulli, W. W. Hargrove, Geospatiotemporal data mining in an early warning system for forest threats in the United States, in: Proceedings of the 2010 IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2010), 2010, pp. 170–173, invited. doi:10.1109/IGARSS.2010.5653935.

[3] R. T. Mills, F. M. Hoffman, J. Kumar, W. W. Hargrove, Cluster analysis-based approaches for geospatiotemporal data mining of massive data sets for identification of forest threats, in: M. Sato, S. Matsuoka, P. M. Sloot, G. D. van Albada, J. Dongarra (Eds.), Proceedings of the International Conference on Computational Science (ICCS 2011), Vol. 4 of Procedia Comput. Sci., Elsevier, Amsterdam, 2011, pp. 1612–1621. doi:10.1016/j.procs.2011.04.174.

[4] J. Kumar, R. T. Mills, F. M. Hoffman, W. W. Hargrove, Parallel *k*-means clustering for quantitative ecoregion delineation using large data sets, in: M. Sato, S. Matsuoka, P. M. Sloot, G. D. van Albada, J. Dongarra (Eds.), Proceedings of the International Conference on Computational Science (ICCS 2011), Vol. 4 of Procedia Comput. Sci., Elsevier, Amsterdam, 2011, pp. 1602–1611. doi:10.1016/j.procs.2011.04.173.

[5] J. Shlens, A tutorial on principal component analysis, Systems Neurobiology Laboratory, University of California at San Diego.
URL `http://www.snl.salk.edu/~shlens/pca.pdf`

[6] P. Alpatov, G. Baker, C. Edwards, J. Gunnels, G. Morrow, J. Overfelt, R. A. van de Geijn, Y. jye J. Wu, Plapack: Parallel linear algebra package (1997).
URL `http://www.cs.utexas.edu/~plapack/`

[7] J. Gunnels, G. Morrow, B. Riviere, R. V. D. Geijny, Plapack: High performance through high level abstraction, in: Proceedings of ICPP98, 1998.

[8] J. Demmel, L. Grigori, M. Hoemmen, J. Langou, Communication-optimal parallel and sequential QR and LU factorizations, arXiv preprint arXiv:0808.2664.

[9] W. W. Hargrove, F. M. Hoffman, Potential of multivariate quantitative methods for delineation and visualization of ecoregions, Environ. Manage. 34 (5) (2004) s39–s60, doi:10.1007/s00267-003-1084-0.

[10] M. ling Shyu, S. ching Chen, K. Sarinnapakorn, L. Chang, A novel anomaly detection scheme based on principal component classifier, in: in Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the Third IEEE International Conference on Data Mining (ICDM03, 2003, pp. 172–179.

[11] A. Lakhina, M. Crovella, C. Diot, Diagnosing network-wide traffic anomalies, in: In ACM SIGCOMM, 2004, pp. 219–230.

[12] D. Schimel, W. Hargrove, F. Hoffman, J. McMahon, NEON: A hierarchically designed national ecological network, Front. Ecol. Environ. 5 (2) (2007) 59. doi:10.1890/1540-9295(2007)5[59:NAHDNE]2.0.CO;2.

[13] M. Keller, D. Schimel, W. Hargrove, F. Hoffman, A continental strategy for the National Ecological Observatory Network, Front. Ecol. Environ. 6 (5) (2008) 282–284, Special Issue on Continental-Scale Ecology. doi:10.1890/1540-9295(2008)6[282:ACSFTN]2.0.CO;2.